LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models

Year: 2023 | *Citations: 165* | *Authors: Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin*

Abstract

Large language models (LLMs) have been applied in various applications due to their astonishing capabilities. With advancements in technologies such as chain-of-thought (CoT) prompting and in-context learning (ICL), the prompts fed to LLMs are becoming increasingly lengthy, even exceeding tens of thousands of tokens. To accelerate model inference and reduce cost, this paper presents LLMLingua, a coarse-to-fine prompt compression method that involves a budget controller to maintain semantic integrity under high compression ratios, a token-level iterative compression algorithm to better model the interdependence between compressed contents, and an instruction tuning based method for distribution alignment between language models. We conduct experiments and analysis over four datasets from different scenarios, i.e., GSM8K, BBH, ShareGPT, and Arxiv-March23; showing that the proposed approach yields state-of-the-art performance and allows for up to 20x compression with little performance loss. Our code is available at https://aka.ms/LLMLingua.